DOCUMENT RESUME

ED 430 021 TM 029 734

AUTHOR Tapper, Rich
TITLE The Problem of High Stakes Assessment in Public Education.
PUB DATE 1997-00-00
NOTE 17p.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Accountability; *Educational Assessment; Educational Policy; Elementary Secondary Education; *Evaluation Methods; Program Development; *Public Schools; *Test Use; Testing Problems
IDENTIFIERS *High Stakes Tests

ABSTRACT

Whether increasing reliance on policy-driven assessment for accountability and control of educational institutions is actually sabotaging long-term goals and purposes of the schools is explored, questioning whether current practices of high-stakes testing are anathema to real education values. The distinction between policy-driven assessment and instructional evaluation is described. The assumptions and purposes underlying scientific and political evaluation as opposed to those of diagnostic assessment are probed, and the work of Jennie Oakes and others is used as the basis for deriving a recommendation for valid, reliable, and appropriate assessments on both individual and institutional levels to facilitate the development of effective schools. Despite the criticism of high-stakes testing, it is not recommended that policy-driven high-stakes tests be abolished. Instead, their rational, effective, and judicious use should be the objective. Formative diagnostic methods and approaches are needed as an integral part of effective instructional programs and program development. (Contains 18 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 430 021

The Problem of High-Stakes Assessment in Public Education

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Richard Tapper

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM029734

Rich Tapper
EDFO 206/Kin White
Spring 1997

BEST COPY AVAILABLE

Rich Tapper
EDFO 206
Spring 1997

May 4, 1998

Rich Tapper
EDFO 206
Spring 1997

In recent years, as it has become necessary to develop large-scale determinants of effectiveness and competence in mass and public schooling, the gaze of educators and educational researchers has increasingly been focused on test reliability and validity. In direct descent from the field of scientific measurement, educational assessment has adopted the values, standards and logic of the scientific method. Particularly of interest to researchers has been the relationship among test design, implementation and interpretation (Messick, 1989). Intuitively and reasonably, researchers and pundits have recognized that these three aspects of valid and reliable assessments must parallel in order for our tests to have power and meaning. What have remained largely unexamined, however, are the crucial distinctions between scientific measurement and educational evaluation – as we have blithely accepted scientific methods into our field. While attention has been directed toward the validity and reliability of our constructs and their uses as decontextualized, scientific measures of performance and achievement, the attendant effects on and affects of our subjects being tested in the name of science and education have been largely ignored.

In this paper we will examine the consequences of such ignorance: through our increasing reliance on policy-driven assessment for accountability and control of our educational institutions, are we unwittingly sabotaging our long-term goals and purposes of our schools? More than this, our issue may be of the nature and current trends of system evaluation: are current practices of high-stakes testing anathema to the values we seek to instill and entrain in our students and encourage in our institutions? This paper will encompass related and integral issues: the distinction between policy-driven assessment and instructional evaluation; the assumptions and purposes underlying scientific and political evaluation as opposed to those of diagnostic assessment; and finally, following the work of Jeannie Oakes, among others, a recommendation for valid, reliable, and appropriate assessments on both the institutional and individual level that facilitate, instead of inhibit, the development of effective schools.

High-Stakes Assessment and Formative Evaluation

In this work, we will refer to large-scale assessments in a number of ways, depending on the context of our discussion. Following Peter Airasian's (1993) work, "high-stakes" testing,

competency-based assessment, measurement-driven instruction and policy-driven assessments will all refer to the practice of large-scale summative evaluation. Such assessment is generally intended to facilitate and catalyze individual and organizational effort, reward satisfactory or exemplary performance, and control curriculum. Although these terms at times connote somewhat different form, implementation and purpose, they have certain tendencies and assumptions in common which make them interchangeable for our analytical purposes.

Large-scale policy-driven evaluations are largely distinguished from instructionally relevant, formative evaluations by their very nature. Policy-driven measurements are, without exception, imposed upon students, schools, and school systems. Such summative evaluations are not concerned with exploring the deep-processing and ultimately meaningful understanding of the examinee, but rather by their very nature examine more easily quantifiable and generalizable constructs. Although in recent years effort increasingly has been applied in developing authentic and direct assessments, instead of computerized multiple-choice exams, as of yet they are impractical to implement on the large-scale. Since the ostensible and popular policy of implementing large-scale exams is to separate the minimally competent from the not quite so, more rigorous examination of the dynamics of learning is perhaps tangential to their purpose. As well, the standards for the assessment are determined at the State or administrative level and so, by implication, few or no allowances can or will be made for local discretion in instruments or their use. Context is, by implication, an irrelevant variable.

As matters of policy and intent, such evaluations usually include moral prescriptions and assumptions, as well. In general, they are designed and implemented as part of a larger scheme to make students or teachers work harder, or reward effort and results in instructional programs. Most importantly perhaps, there are important consequences, or "high stakes" associated with determined achievement or performance levels. More often than not, consequences are public ones. This last characteristic in itself may contribute greatly to the practical difficulty with many high-stakes exams as schools and systems strive to look good, generally before other concerns.

Although such summative evaluation in design is oftentimes at odds with our formative, instructional purposes, it is one of the few easily and popularly quantifiable handholds on which policy-makers and administrators can grasp. Fuzzier concepts, while perhaps more central to

development and education, are much more difficult and costly to measure. In fact, they may be wide of the point if our purpose, quite simply, is to check for minimum levels of product.

In contrast, as Gagne and Bloom among others have explored in their instructional taxonomies, *formative* evaluation and feedback are an integral part of any effective educational program, as a homeostat is a necessary part of a developing organism or effective heating system; it is a component of instructional events leading to effective mastery in almost any domain. Without formative evaluation and feedback, our students are left to reach end-states like ships without rudders, sailing without navigation. Without proper evaluation, our students, or we cannot be sure that we are moving in the appropriate instructional or methodological direction. This is true at both the micro and macro level; it is true for both individuals and organizations. The crucial characteristic of such evaluation is that it is formative and diagnostic – it is oriented toward the process of learning and development. Even more than a unidimensional evaluation, however, such assessment is known to have effects on the learner. There is an inherent understanding that the subject of our analysis is a participating actor in the process, and so effected, in both achievement and motivation for future achievement, through and by the process.

The distinction between these two types of assessment is clearly in intention and purpose: in a logical and rational world, design of the evaluations should follow. Further, as Stiggins (1993) notes, centralized assessment and classroom (instructional) assessment differ in more than just scope, and so should differ in appropriate use and possible consequences for the student and system. Relevant to our discussion is an understanding that the roles of teachers and policy-makers seeking data are quite dissimilar. Clearly and often policy-makers set as their goal to attain the highest possible scores, while teachers seek accurate accounts of their students' strengths and weaknesses to help meet their needs. Policy-makers and scientists measuring static outcomes try to eliminate or minimize "standard error" in their tools, while for teachers, this discrepancy or variation among students is one of the most essential for addressing student needs – such diagnostic information was, in fact, one of the catalysts for Piaget's theory of development. For policy-makers, summation is the key to effective management, while teachers seek formative evaluation in their understanding of the process of the dynamic interaction with their students. For policy-makers, quality large-scale, high-stakes assessments "are seen as the guardians of our educational standards," (Stiggens, p. 96), while for the teacher, quality

assessments are not (technical) matters of reliability and validity, but rather matters of the impact on student development and motivation. These differences take us to the doorstep of the distinction between large-scale and classroom assessment, between policy-driven and instruction-driven evaluation.

The Purpose of Assessments

In contrast to policy-driven assessment, which leads only to high-stakes consequences for examinees or systems, formative and/or diagnostic evaluation helps students and systems identify the source of their errors and the cause of their errors – and so lead to correction. Appropriate formative evaluation, including appropriately communicated feedback for the student or system, also tends to facilitate the development of intrinsic motivation, while high-stakes assessment with its emphasis on static product and consequences, tends to *inhibit* motivation. In the recent motivation and attribution literature (Deci, 1992), as well as classic research on parenting styles (Baurmrind, 1991; Dornbusch, et al, 1987; Lamborn, et al, 1991), it is clear that controlling, authoritarian contexts are contraindicated for healthy development of mind and psyche. As Deci has examined regarding the relationship of evaluation to educational outcomes, "when people are motivated by control or pressure…intrinsic motivation and interest that students have for learning tends to be undermined. This, in turn, impairs their conceptual understanding of the material." (p. 63)

The crucial distinction here is in the purpose and intention of our assessments, both policy-driven and instructional, evoking the logic of Messick (1989): "The essence of unified validity is that the appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable..." (p. 5) Intuitively, we know this to be true, and yet there is still a popular call for accountability in the form of consistently implemented high-stakes, mandated testing to ensure the success of our schools – in short, for policy-driven, summative evaluation to function as diagnostic homeostat, albeit poorly and counterproductively.

The implied values of high-stakes assessments mean little or nothing when used for formative purposes. Further, they tend to sabotage the instructional and developmental process when used as such, for they are indeed controlling by their very nature and intent. Such logic as is used by policy-makers to support their decisions to implement such overtly controlling

measures is similar to the rationale of controlling and authoritarian parents, unable or unwilling to act in a ways more appropriate to facilitate the autonomy and healthy development of their children. The dysfunction resulting on the level of the family clearly has its analogy in the organization.

As we have seen, concerns and uses of policy-driven and instructional assessments are divergent, and yet these two ends of our evaluative spectrum are often conflated when large-scale assessments are used for anything but threshold determinants of minimal effectiveness. Policy-driven evaluation as applied to education is most appropriately a threshold measure, particularly of the superficial reflections of understanding currently acting as the educational currency most easily comprehended by policy-makers. It can be a powerful tool in our quest to ensure a minimum level of competence and development for our students, but as in the old saw about a powerful medication, however, a few drops will cure, while an ounce may kill. For example of unwise and over-use, we have only to turn to the current predominance of End of Grade tests (EOGs), and the spate of high-stakes assessments in states like New York, used to monitor school effectiveness and so dictate and control curricula and policy. In states like North Carolina, ABC models of policy lead to direct (punitive) control of curricula by the state.

Current Practices of Classroom Assessment

A dark reflection of this mentality is in classrooms across the nation, in which teachers untrained and unskilled in test theory, and ignorant of the constructs they are measuring, design tests which facilitate nothing but a puerile fascination with peripheral detail. The educator Jacques Barzun bemoans their effects in an essay from *Begin Here: The Forgotten Conditions of Teaching and Learning*,

> Because the modern world lives by machine industry, it favors the mechanical in all things, whether all things benefit from it or not. We judge of the known and the unknown by numbers and make do even with indirect clues to them – so-called indicators…The answers are totted up according to a code, and on the basis of it the hiring is done or the prescription written…That numerical remote control has invaded the school in the form of multiple-choice tests, and their obvious convenience has concealed a series of harmful side-effects…on the minds of the

> learners and on the meaning of things taught...[There are] equally bad consequences for other prime elements of schooling. (p. 28)

The very qualities that make a make a test generalizable and reliable make it anathema to the development of diversity and true intelligence; they make it impossible for our students to follow the advice of Emerson: "Tell us what you know." Barzun makes the case for essay examinations and other tests of creative recall and construction as true and appropriate evaluations if we are to emphasize deep learning and develop the minds of our students.

MacIver and Reuman (1994) make a similar case for the use of grading and recognition practices that motivate students to work hard, and a case against those that do not: "Traditional assessment, grading, and student recognition practices are partly responsible for the anti-academic norms and low levels of student effort that pervade American schools." The authors go on to cite programs inspired by an understanding of motivation and appropriate accountability, resulting in significant improvements in student attitude, peer support, and overall achievement. Such programs are, however, rare. As Howard Gardner has so pithily expressed it in *The Unschooled Mind*:

> Even though educational systems may pay lip service to goals like "understanding" or "deep knowledge," they in fact prove inimical to the pursuit of these goals. Sometimes these goals are considered to be hopelessly idealistic or unrealistic; at most, in the view of educational bureaucrats, schools ought to produce citizens who exhibit some basic literacies and can hold a job. But even in cases where these goals are taken seriously, events conspire to undermine their pursuit. Particularly when systems are expected to produce hard evidence of their success, the focus sooner or later comes to fall on indices that are readily quantified, such as scores on objective tests. Measures of understanding must be postponed for another day or restricted to a few experimental schools, which are allowed to operate under waivers. (p. 140)

Implications of Wrong Use

By implication then, large-scale, high-stakes assessments should not and cannot be legitimately used as formative, instructional feedback, for either an individual or system, but

rather as checks on minimal (and oftentimes superficial) competence. As explained by the US Congress' Office of Technology Assessment, the current trend toward such policy-driven controls on education are an outgrowth of a reactionary back-to-basics movement, largely spearheaded by those outside the art of education, and ignorant of the difference between process and product, and the crucial determinant of school context. Growing from a widespread assumption that more control is necessary to head-off the deterioration of our educational system (perhaps fueled by schools without walls, the experimental education movements of the early 1970s, and the popular notion that declining SAT scores reflect a state of general educational malaise), the movement was further fed by the tendency of States to pick up a larger share of the educational tab. Paying the piper, policy-makers clearly decided that the tune to be played was accountability in the same key as that played for licensure exams in professional fields, already widespread and widely accepted for many years. As in the old Sufi story, the light may be better in the areas which are easily quantified and compared to justify our policies, but the focus is in decidedly the wrong place if we truly wish to understand and effect positive change.

The implications of the control by the State through evaluation are self-evident. As well, the attendant organizational and motivational assumptions on which we base such educational policies our suspect if not downright faulty. The popular use of large-scale assessments as methods of systemic evaluation springs from a "fundamental notion that if people are to be judged according to certain types of criteria, they will try to excel with respect to those criteria." (Popham, p. 32) While this is intuitively attractive, it is flawed in the practice of confusing the static map with the dynamic territory, in confusing product with process, the peripheral with the crucial, the scientific with the educationally meaningful. What is easily quantifiable in education is not the essence of what we seek to develop, but only its reflection. Like the inhabitants of the cave in Plato's *Republic*, we begin to hold the shadow in the highest esteem, and disregard the world of substance in our choice of curriculum and method.

More often than not, such diversion from what is central to education and development results in detrimental effects for both students and systems. Since such large-scale summative measures, by their very nature, can and do measure only the most basic and minimum competencies, they serve as a magnet for only the most superficial of constructs, however difficult it may be pass. As quoted by Airasian, Popham (1987) explains:

> Measurement-driven instruction occurs when a high stakes test of educational achievement, because of the important contingencies associated with the students' performance, influences the instructional program that prepares students for the test...Teachers tend to focus a significant portion of their instructional activities on the knowledge and skills assessed by such tests. A high stakes test of educational achievement, then, serves as a powerful *curricular magnet* (p. 680)

As it is on the level of the individual, such high-stakes assessments tend to stifle the intrinsic motivation of organizations for the deep and meaningful, correspondingly influencing the curriculum and instruction. Such evaluation directs individuals and systems toward the goal of passing the test, of achieving the set standard. More often than not, satisfactory passing or achievement levels on these superficial peripheral constructs are rewarded through carrots on sticks, such as bonus structures for administrators, preferred treatment for systems, or avoidance of punitive consequences, and so reinforce vacuous and self-serving approaches to education. Clearly, it is in the best interest of the school as a component organism of the State to maximize student achievement levels on such summative exams, regardless (and often in spite) of unforeseen and often deleterious effects on student motivation and overall development.

Ironically, such large-scale assessments rely on the dubious assumption that increased accountability and proscription will translate to greater achievement and effective schooling – the assumptions of scientific management applied to education:

> Although [the] ability of a Statewide testing program to control local activity may be praiseworthy in the minds of some educational critics, the activity... stimulated [is] not reform. Responding to testing... [does] not encourage educators to reconsider the purposes of schooling; their purpose quickly ... [becomes] to raise scores and lower the pressure directed toward them. (Corbett and Wilson, 1990, p. 10-11)

The rationale of policy-makers in this regard is evidently suspect. To focus our evaluative attention on such objectives is generally to neglect our other concerns, particularly when

consequences are high. George Madaus, director of the Center for the Study of Testing, Evaluation, and Educational Policy expressed the problem of the influence of such assessment:

> When the stakes are high, people are going to find ways to have test scores go up...The school will look better, but the skill levels will not necessarily be going up. You may have succeeded only in corrupting the inferences you wanted to make from the tests (Allington and McGill-Franzen, p. 3)

Allington and McGill-Franzen (1992) explored this very notion in their research on the effects of high-stakes testing on New York public schools.

The Effects of High-Stakes Testing

Since 1985, New York has instituted a series of high-stakes assessments in order to effect an improvement in the reading levels of students by the identification of children with extra educational needs. The reading tests are designed to: 1) target children at risk of school failure so that they can receive instructional support, and, 2) become a part of the public accountability profiles compiled annually by the State Education Department. In sum, Allington and McGill-Franzen found that rising scores within individual schools and systems was due not to improvements in reading or reading instruction, but rather the breakup of cohorts as students were retained in grade or (intentionally or unintentionally) left out of annual testings by individual schools. Clearly, the summative reading evaluations presented in the profiles of the examined systems were far more subject to interpretation than publicly acknowledged or understood by policy-makers.

Further, the authors found that tactics used to raise scores had very little correlation with effective strategies to improve reading, one of the consistent pitfalls of summative, high-stakes assessment. Such testing may, in fact, act counter to intentions to facilitate deep learning and more effective developmental programs – in the same way that multiple-choice exams act counter to this same intention in the individual student. In this particular study, schools reporting the highest percentages of students passing the third-grade competency exam also made the heaviest use of questionable instructional practices such as retention in grade and special education placement. "The achievement of children after they have been identified as mildly

handicapped and placed in special education is disappointing…and as Shepard and Smith (1990) have examined…the consistently dismal record of retention in grade suggests that this practice may serve the needs of schools rather than the needs of children." (Allington and McGill-Franzen, p. 11)

Rather than addressing crucial, contextual aspects of the effectiveness of programs, such testings rely on the decontextualized and scientifically quantified. In a field such as education, and concerning the dynamic development of children, such emphasis may be inimical to the true effectiveness of our schools. Further, the validity of these tests, following from their use, depends upon the assumption that the passing and achievement rates are accurately reflective of the school's instructional program. As we have seen, this latter assumption is a confusion of measurement and construct: there is little or no correspondence between test constructs and instructional programs, although the tests themselves may be valid and reliable. The variable represented is rather the smoke from the fire – a useful gauge signal, but clearly not the heart of the matter. We may be able to tell that there is, indeed a fire, but other than that we are at a loss to describe it, even if our instruments are very, very good at distinguishing different kinds of smoke. As well, the fire may have been put out some time ago – but still the smoke lingers, and the trees smolder. Other fires give off very little smoke, and so will never be detected at all.

A Recommendation

Appropriate methods of formative and diagnostic assessment must be accurately implemented with an awareness nature of education if our intention is to promote the effectiveness of our systems and the development of our students. As Jeannie Oakes has examined, following the work of Lev Vygotsky, education is a dynamic, contextually robust enterprise. Oakes (1989) makes the argument that valid indicators would and should include assessments of school context, in much the same way that appropriate and prompt feedback is a necessary part of instructional events leading to achievement and development. As Oakes has observed, "such information is essential if [we] want monitoring and accountability systems to mirror the condition of education accurately or to be useful in making improvements." (p. 182) Further, she states, "if policy makers choose not to monitor context, they will fail to recognize that school characteristics mediate the effect of educational inputs…Doing so, they will create

monitoring systems that provide a superficial and simplistic portrayal of the educational system." (p. 183) In effect, such monitoring systems encourage schools to marshal a considerable effort in order to look good for the process of evaluation, and by implication neglect what is crucial for education.

Oakes makes the case for three context indicators as descriptors about central features of the educational system: access to knowledge, press for achievement and professional teaching conditions. Following both logic and research, these indicators give a more complete picture of the performance of the education system. Though they focus on less tangible and quantifiable aspects of systems, they are the *alterable characteristics* (her italics) crucial for school improvement, and so the legitimate focus for formative evaluation intended to effect positive change.

Appropriate and Meaningful Assessment

The recommendation made here is not for the abolition of policy-driven, high-stakes tests, but rather for their rational, effective and judicious use. If we seek diagnostic information about our systems in order to improve them, then clearly we ought to embrace formative, diagnostic methods and approaches – including all relevant contextual information, since this is really the heart of the matter. Formative evaluation is an integral part of effective instructional programs and program development; it is more than a scientific measurement of achievement.

By their nature, high-stakes assessments to ensure minimum competency should only be used sparingly, and only as threshold measures. On the organizational front, it is becoming clearer that the "traditional" notions of hierarchical rule function more aptly to repress individual effort than reward and encourage it. In the sphere of education, particularly, it is becoming most evident that those organizational structures (what Jonathan Kozol terms German model of efficiency and scientific management), which tend to most successfully facilitate the efficient production of undifferentiated widgets tend to least successfully develop the idiosyncratic miracles of individual young minds and bodies.

The question before us is how to evaluate our schools so that our societal commitment to education is not hindered but rather expanded and empowered. Giving power and responsibility back to the local systems with an understanding of the crucial role of context and process is at the

root of our strategy. At the same time, there must be legitimate provisions to ensure a minimum quality of education in places economically, intellectually, or materially impoverished. Such a concern, mixed with a misunderstanding of the effects of testing and control, provided the background for "traditional" solutions to problems of educational policy, including federal mandates of curriculum (in effect) thrust on states and localities.

The apparent dissonance between necessary federal and State guarantees of minimum competence and the educational necessity for local dictate of method fuels our debate. Clearly, any effort to reduce the role of the State in education runs the risk of disenfranchising a significant minority dependent upon the sponsorship and protection of State policies ensuring minimum competency. Can we deconstruct our nation's responsibility to educate its citizens into aims and goals with which <u>every</u> state would agree without compromising self-determination of method and aims unique to each province and locality? This would seem to be our starting point in developing appropriate policy-driven, high-stakes assessments.

I would suggest that our overriding and common concerns as a nation include not aggregate, quantifiable achievement in science and math, nor even graduation rate from and seat-time in secondary school, but rather *literacy* and democratic viability for all our citizenry. As such, our strategy to reduce our reliance on and compliance with federal mandates of curriculum and method, while maintaining a national standard of agreement about minimum competency standards, is to reduce all nationwide, normed assessments of student achievement to simple literacy and critical thought: reading, writing and 'rithmetic. Assessments of reading comprehension and critical reading might look like SAT problems, though much more exhaustive. Assessments of writing and critical thinking might take the form of essays, judged by raters trained through national standards, obviating locally prejudiced standards and "grade inflation". Such evaluations might occur at the end of elementary, middle and secondary school, culminating in a national exam for graduation.

We as a nation might promote local responsibility through the same tactic used by many parents to wean their children: trust our localities to do what is in the best interests of our nation as a community, and *verify* that our trust is well-placed through reports of results and methods. In addition to sparingly used high-stakes assessments as threshold indicators, such verification might include quarterly, semi-quarterly or annual reports to state commissions vested with the

power to challenge the methods of localities. Reports, as such, would include school mission-statements and attendant educational goals, methods and practice – an *authentic* portfolio assessment of an entire school, much in line with the recommendations and research of Jeannie Oakes.

Such a strategy to promote local control and responsibility might also include a shift of focus of public discourse, from graduation percentage and nationwide scores on tests of math and science to a conversation about what truly makes a democracy a democracy: the ability of its citizens to *think*. Clearly, it is the responsibility of the State to ensure full compliance through oversight – particularly when State monies are being spent. Just as clearly, each locality ought to have the power to determine their own context. As in politics, all education is local. As our policies and expectations shift, so will the results of our schools. Our current educational crisis is an opportunity to refocus our attention on what is truly important in education – not evaluations on easily quantifiable constructs, but the interactions present in the microcosm of the student-teacher-school relationship, in our localities as *learning communities*. Our choice of evaluations can and should reflect and facilitate this understanding, or we are truly compromising the effectiveness of our schools and the development of our students.

References

Airasian, Peter. (1993). Policy-driven assessment or assessment-driven policy? *Measurement and Evaluation in Counseling and Development,* vol. 26 April 1993, 22-30.

Allington, R. L., & A. McGill-Franzen. (1992). Does high-stakes testing improve school effectiveness? *ERS Spectrum, 10*(2), 3-12.

Barzun, Jacques. (1991). Middlemarch *Is*---? Reason to de-test the schools. In Morris Philipson (Ed.), *Begin Here: The Forgotten Conditions of Teaching and Learning.* (pp. 28-32). Chicago and London: The University of Chicago Press.

Baumrind, D. (1991). The influence of parenting style on adolescent competence and substance abuse. *Journal of Early Adolescence, 11,* 56-94.

Clifford, Margaret M., et al. (1990). Academic risk taking, development, and external constraint. *Journal of Experimental Education, 59*(1), 45-64.

Corbett, H. D., & Wilson, B. (1990). Unintended and unwelcome: The local impact of state testing. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA, April 1990.

Deci, Edward L. (1992). The relation of interest to the motivation of behavior: A self-determination theory perspective. In K.A. Renninger, S. Hidi, and A. Krapp (Eds), *The Role of Interest in Learning and Development* (pp. 43-70). Hillsdale, NJ: Lawrence Earlbaum Associates, publishers.

Dornbusch, S. M., Ritter, P., Leiderman, P. H., Roberts, D. F., & Fraleigh, M. J. (1987). The relation of parenting style to adolescent school performance. *Child Development, 58,*, 1244-1257.

Gardner, Howard. (1991). *The Unschooled Mind: How Children Think and How Schools Should Teach.* New York: Basic Books.

Lamborn, S. D., Mounts, N. S., Steinber, L., & Dornbusch, S. M. (1991). Patterns of competence and adjustment among adolescents from authoritative, authoritarian, indulgent and neglectful families. *Child Development, 62,* 1049-1065.

MacIver, Douglas, & Reuman, David A. (1993). Giving their best: Grading and recognition practices that motivate students to work hard. *American Educator, 17*(4), 24-31.

Messick, Samuel. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher,23*(2), 13-23.

Messick, Samuel. (1989). Meaning and values in test validation: the science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Oakes, Jeannie. (1989). What educational indicators? The case for assessing the school context. *Education Evaluation and Policy Analysis, 11*(2), 181-199.

Popham, James W. (1993). Measurement-driven instruction as a "quick-fix" reform strategy. *Measurement and Evaluation in Counseling and Development,* vol. 26 April 1993, 31-34.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan. 68*(9), 679-682.

Stiggins, Richard J. (1993). Two disciplines of educational assessment. *Measurement and Evaluation in Counseling and Development,* vol. 26 April 1993, 93-104.

US Congress. Office of Technology Assessment. (1992). *Testing in American School: Asking the Right Questions.* Washington, DC: US Government Printing Office.